MDPI

Article

Gráfico de control T^2 Hotelling para variables cualitativas

Wilson Rojas-Preciado[1,2], Mauricio J. Rojas-Campuzano[3], Purificación Galindo-Villardón[2], Omar Ruiz-Barzola[3], , ,

* Correspondence: wrojas@utmachala.edu.ec; Tel.: +593-992-83-3719
† Current address: Updated affiliation
‡ These authors contributed equally to this work.

Version July 18, 2022 submitted to Water

Simple Summary: A Simple summary goes here.

Abstract: La literatura científica es abundante en lo referente a gráficos de control en entornos multivariantes para datos numéricos y mixtos, sin embargo, para datos cualitativos hay pocas publicaciones. Las variables cualitativas aportan valiosa información de procesos en diversos contextos industriales, productivos, sociales. Los procesos educativos no son una excepción, tienen múltiples variables asociadas a estudiantes, profesores e instituciones. Cuando hay muchas variables se corre el riesgo de tomar información redundante o excesiva, luego, es viable la aplicación de métodos multivariantes de reducción de dimensiones para quedarse con pocas variables ficticias, combinación de las reales, que sinteticen la mayor parte de la información. En este contexto se presenta el gráfico de control T2Qv, una técnica de control estadístico de procesos multivariantes que realiza un análisis de datos cualitativos mediante Análisis de correspondencias múltiples (MCA), Análisis Factorial Múltiple y el gráfico T^2 de Hotelling. La interpretación de los puntos fuera de control se realiza comparando los gráficos MCA y analizando la distancia X^2 entre las categorías de la tabla concatenada y las que representan puntos fuera de control. El análisis de sensibilidad determinó que el gráfico de control T2Qv tiene un buen rendimiento cuando trabaja con altas dimensiones. Para probar la metodología se hizo un análisis con datos simulados y otro con datos reales relacionados con la educación superior. Para facilitar la difusión y aplicación de la propuesta, se desarrolló un paquete computacional reproducible en R, denominado T2Qv y disponible en The Comprehensive R Archive Network (CRAN).

Keywords: keyword 1; keyword 2; keyword 3 (list three to ten pertinent keywords specific to the article, yet reasonably common within the subject discipline.).

1. Introduction

Los gráficos de control constituyen una de las herramientas más importantes para definir límites y parámetros óptimos de los procesos, así como para controlar la calidad de los productos mediante la reducción de la variabilidad. El uso de gráficos de control facilita la evaluación del comportamiento de las variables del proceso y contribuye al logro de los objetivos planificados.

La variación de los procesos se entiende como la diversidad de resultados que genera un grupo de variables de un proceso, su monitoreo es un objetivo clave del control estadístico, por lo tanto, es necesario entender los tipos y motivos de la variabilidad. Para ello es preciso registrar de manera sistemática y adecuada diferentes variables del proceso que se desea controlar, como las propiedades

de los insumos, las condiciones de operación de los equipos, las competencias del personal que maneja los procesos, además de las características de los productos, la satisfacción de los usuarios, el cumplimiento de requisitos, entre otras.

El pionero del control estadístico de procesos fue Walter Shewhart (SPC). Estableció las diferencias entre la variabilidad natural o común, presente en todos los procesos, y la provocada por causas asignables o especiales, que pueden llevarlos a un estado de fuera de control. Señaló que un proceso está en control estadístico cuando trabaja sólo con causas comunes de variación. Propuso los primeros gráficos de control para variables de tipo continuo y para variables de atributos [1].

El SPC mediante gráficos de control permitió a las organizaciones monitorear el comportamiento de una variable a la vez, no obstante, las organizaciones requirieron, con el pasar del tiempo, el análisis de varias características de calidad de forma simultánea, abriendo la puerta al SPC desde una perspectiva multivariante [2]. Para facilitar el control de calidad de procesos es común el uso de gráficas de control que recolectan abundante información en diversas variables de forma simultánea, su análisis permite caracterizar los diferentes tipos de variables que afectan la calidad y explican su comportamiento a lo largo del tiempo [3].

Hay una variedad de gráficos de control de procesos desde la perspectiva multivariante, entre los clásicos están el Gráfico T^2 de Hotelling [4], el Multivariate Exponentially Weighted Moving – MEWMA [5], el Multivariate Cumulative Sum Control Chart – MCUSUM [6]. Con el transcurso del tiempo se hicieron diversos aportes para mejorar el rendimiento de estos gráficos, entre los más destacados están Gráfico de control T^2 con tamaños de muestra adaptables [7], Gráfico de control T^2 con intervalos de muestreo variables [8], Gráfico de control T^2 con líneas de advertencia dobles [9], Gráfico de control robusto [10], Gráficos de control basados en modelos de minería de datos para procesos multivariantes y autocorrelacionados [11], Gráficos de control de calidad multivariantes con dimensión variable [12], Gráfico de control para el coeficiente de variación multivariante [13].

Además de estos gráficos de control para entornos paramétricos, se desarrollaron otros para datos numéricos y cualitativos en entornos multivariantes no paramétricos, entre ellos el Gráfico de control multivariante basado en la distancia de Gower para una combinación de variables continuas y cualitativas [14], Gráfico de control multivariante basado en la combinación de PCA para características de calidad de atributos y variables [15], Gráfico de control multivariante no paramétrico basado en la ponderación de novedad sensible a la densidad para procesos no normales [16], Gráfico de control de deméritos con clustering difuso de c-medias [17], Gráfico de control basado en ACP que utiliza máquinas de vectores de soporte para distribuciones no normales multivariadas [18], Gráfico CUSUM no paramétrico para monitorear procesos multivariados correlacionados en serie [19], Gráfico de control multivariante basado en Kernel PCA para monitorear características de calidad de atributos y variables mixtas [20], Gráfico T^2 basado en la combinación de PCA para datos continuos y cualitativos con detección de datos atípicos [21].

Como se puede observar, la literatura científica es abundante en lo referente a gráficos de control en entornos multivariantes paramétricos y no paramétricos para datos numéricos y, en los últimos años, para datos mixtos (numéricos y cualitativos), sin embargo, son pocas las publicaciones sobre gráficos de control multivariantes para datos cualitativos. En este campo las propuestas se han desarrollado alrededor del análisis de variables que siguen una distribución Poisson y el análisis de variables multinomiales.

La primera propuesta fue la de Holgate [22], quien presentó un trabajo sobre la distribución Poisson bivariante para variables correlacionadas. Este modelo fue tomado como insumo en las investigaciones de autores como Chiu and Kuo [23], Lee and Costa [24], Laungrungrong *et al.* [25], Epprecht *et al.* [26]. Otra propuesta destacada es la de Lu [27], quien desarrolló un gráfico de control tipo Shewhart para procesos multivariados con variables de atributos, cuando la característica de calidad asume valores binarios, que se denominó gráfico *np* multivariante (MNP). No obstante, hay escenarios en los que una clasificación dicotómica es insuficiente y se vuelve necesario acudir a niveles intermedios, en cuyo caso el análisis requiere el uso de distribuciones multinomiales.

En este contexto Mukhopadhyay [28] planteó un gráfico de control multivariante utilizando el estadístico D^2 de Mahalanobis para atributos que siguen una distribución multinomial. Además, surgieron los gráficos de control multivariantes en procesos multinomiales bajo el enfoque difuso [29]; Taleb [30] introdujo gráficos de control para el monitoreo de procesos multivariados con datos lingüísticos multidimensionales, basados en dos procedimientos: la teoría de la probabilidad y la teoría difusa; Fernández *et al.* [31] presentaron un gráfico de control multivariante multinomial T2 con un enfoque difuso.

Un aporte interesante es el de Epprecht *et al.* [26], quienes presentaron una combinación lineal óptima de variables discretas, cuando siguen la distribución de Poisson, para el SPC multivariantes. Asimismo, Ali and Aslam [32] desarrollaron gráficos de control para datos con distribución Poisson multivariante utilizando un muestreo generalizado de estados dependientes múltiples (GMDS).

Saltos Segura *et al.* [33] aseguran que las herramientas de control de la calidad se pueden considerar no solo para monitorizar procesos industriales sino también procesos relacionados con la educación, por ejemplo, la evaluación del desempeño estudiantil. Estos autores aplicaron el concepto de profundidad, que transforma una observación multivariante a un índice univariante, el cual es susceptible de monitorizar en una carta de control y para esto utilizaron la carta r, además utilizaron clúster medio para establecer umbrales que faciliten la conformación de grupos y establecer perfiles de estudiantes mediante medidas descriptivas.

En el estudio de los procesos que se desarrollan en el entorno social-educativo y que explican el comportamiento de variables como el rendimiento académico, tasas de graduación o deserción, producción científica, porcentajes de matrícula de nuevo ingreso, entre otros, se maneja con mucha frecuencia variables cualitativas. No es que estén ausentes los datos cuantitativos, sino que, en las bases de datos que se utilizan para estos análisis, abundan las variables cualitativas nominales y ordinales sobre las de tipo numérico. Algunos ejemplos de datos de los estudiantes son: sexo, lugar de procedencia, autodenominación étnica, grado académico de los padres, tipo de institución educativa de procedencia (fiscal, particular, municipal), asistencia, seguimiento al sílabo, resultado (aprueba, no aprueba); ejemplos de datos de las instituciones son: tipo de sostenimiento económico, jornada, modalidad, campo de estudio, niveles (tecnológico, grado y postgrado), tipo de infraestructura; ejemplos asociados a datos de los profesores son: titularidad, dedicación, grado académico, grado en el escalafón, discapacidad, nivel de capacitación, avance académico, resultados de la evaluación de desempeño, entre otros.

López [34] señala que al observar muchas variables sobre una muestra es presumible que una parte de la información recogida pueda ser redundante o que sea excesiva, en cuyo caso los métodos multivariantes de reducción de la dimensión tratan de eliminarla combinando muchas variables observadas para quedarse con pocas variables ficticias que, aunque no observadas, sean combinación de las reales y sinteticen la mayor parte de la información contenida en sus datos. En este caso se deberá tener en cuenta el tipo de variables que maneja. Si son variables cuantitativas las técnicas que le permiten este tratamiento pueden ser el Análisis de componentes principales [35,36], el Análisis factorial [37–39], mientras que, si se trata de variables cualitativas, es recomendable la aplicación de un Análisis de correspondencias múltiple, Análisis de homogeneidad o un Análisis de Escalamiento multidimensional.

En el control estadístico de procesos, los aportes al desarrollo de gráficos de control para variables cualitativas todavía son incipientes, las pocas publicaciones se orientan al análisis de características de la calidad en procesos industriales, pero no a procesos sociales como la educación. Al analizar los procedimientos publicados por los autores citados en este estudio, se detectan limitaciones que podrían restringir su aplicación, por ejemplo, el análisis de pocas características de la calidad, el uso de muestras constituidas por elementos individuales en vez de grupos, la dificultad de trabajar con muchas categorías de forma simultánea. Surge, entonces, la necesidad de un gráfico de control para la representación de p variables cualitativas, que pueda trabajar con múltiples categorías nominales y

ordinales, que facilite la identificación de las causas que pueden llevar al proceso a un estado fuera de control y que pueda ser aplicado también procesos sociales.

Esta necesidad se atiende en esta investigación, cuyo objetivo es desarrollar un gráfico de control para variables cualitativas con múltiples categorías nominales y ordinales, mediante la aplicación de una metodología de análisis multivariante, para que se contribuya a la diversificación de técnicas en la fase I del control estadístico de procesos.

Este artículo está organizado de la siguiente manera: la Introducción, que establece los antecedentes conceptuales y referenciales de los gráficos de control multivariantes aplicados a variables cualitativas; la sección 2, materiales y métodos, que detalla el procedimiento que se siguió en el desarrollo del gráfico de control propuesto; la sección 3 describe al complemento computacional que facilita la aplicación de esta metodología; la sección 4 muestra los resultados mediante el análisis de datos simulados y datos reales aplicados al contexto de la educación superior; la sección 5 corresponde al análisis de sensibilidad que relaciona el número de dimensiones analizadas versus la confiabilidad de los resultados. La sección 6 presenta la discusión mediante un análisis comparativo entre el gráfico de control T2Qv y las propuestas de otros autores. Finalmente, la sección 7 establece las conclusiones.

2. Metodología

2.1. Notación

La tabla 1 contiene elementos, representación y ejemplos de la manera como se presentan los elementos algebraicos abordados en la metodología.

Elementos	Representación	Ejemplo
Escalares	Letras en minúscula.	v, λ
Vectores	Letras en minúscula y en negrita.	$\mathbf{v}, \mathbf{u}$
Matrices	Letras en mayúscula y en negrita.	$\mathbf{V}, \mathbf{X}$
Matrices de tres vías (Cubos de datos)	Letras con doble trazo en mayúscula.	$\mathbb{C}, \mathbb{X}$

Table 1. Elementos algebraicos

A lo largo del artículo se utilizarán letras para hacer referencia a parámetros necesarios, se los enuncia a continuación en la tabla 2:

Letra	Significado	Especificación
p	Número de dimensiones	
K	Número total de tablas (Especifica la profundidad del cubo de datos)	
k	Índice de tabla	k=1,2,...,K
T	Índice de matriz transpuesta	$\mathbf{X^T}$
n	Tamaño muestral de las k tablas	

Table 2. Notación

2.2. Bases metodológicas

Para facilitar la explicación de la metodología se ha generado una base de datos que contiene 10 tablas, cada una con 10 variables categóricas, con los niveles *Bajo*, *Medio* y *Alto*, denominada *Datak10Contaminated*, que se describe con detalle en el apartado 4.1.

2.2.1. Análisis de Correspondencias Múltiples (MCA)

El tratamiento multivariante de variables cualitativas requiere un proceso metodológico distinto al que se aplica con variables cuantitativas, uno de los más representativos es el Análisis de

Correspondencias [40]. Según López [34], este análisis implica estudios de similaridad o disimilaridad entre categorías, se debe cuantificar la diferencia o distancia entre ellas sumando las diferencias cuadráticas relativas entre las frecuencias de las distribuciones de las variables analizadas, lo que conduce al concepto de la χ^2. Así, el análisis de correspondencias puede considerase como un análisis de componentes principales aplicado a variables cualitativas que, al no poder utilizar correlaciones, se basa en la distancia no euclídea de laχ^2.

En el enfoque francés del análisis de correspondencias, que se caracteriza por el énfasis en la geometría, el análisis de una tabla cruzada se llama análisis de correspondencias (CA) y el análisis de una colección de matrices indicadoras, se denomina análisis de correspondencias múltiples (MCA) [41]. En contextos anglosajones, el MCA es conocido como Análisis de Homogeneidad o Escalamiento Dual, especialmente en psicometría.

El análisis de correspondencias múltiples (MCA) es una generalización del análisis de correspondencias simple o binario, donde se incluyen más variables cualitativas. Se obtiene al realizar el análisis de correspondencias simple a una tabla disyuntiva completa, conocida como la tabla de Burt.

V_1	V_2	$\cdots$	V_p
Alto	Medio	$\cdots$	Medio
Medio	Bajo	$\cdots$	Alto
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Bajo	Alto	$\cdots$	Bajo

Table 3. Matriz inicial

Esta matriz es equivalente a la matriz disyuntiva Z, que desglosa las variables en cada una de sus modalidades y se registra la ocurrencia de eventos de forma binaria.

V_1 : *Alto*	V_1 : *Medio*	V_1 : *Bajo*	V_2 : *Alto*	V_2 : *Medio*	V_2 : *Bajo*	$\cdots$	V_p : *Alto*	V_p : *Medio*	V_p : *Bajo*
1	0	0	1	0	0	$\cdots$	0	1	0
0	1	0	0	1	0	$\cdots$	1	0	0
0	0	1	0	0	1	$\cdots$	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
1	0	0	1	0	0	$\cdots$	1	0	0

Table 4. Matriz disyuntiva Z

La tabla de Burt viene dada por:

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z} \tag{1}$$

La construcción de la matriz de Burt se da por la superposición de tablas. En las tablas ubicadas en la diagonal se encuentran matrices diagonales que contienen las frecuencias marginales de cada una de las variables. Fuera de la diagonal de la matriz de Burt se encuentran las tablas cruzadas por pares de variables.

Para realizar el análisis de correspondencias múltiples se parte de la matriz de Burt, obtenida con la ecuación 1. Esta matriz está formada por las frecuencias absolutas, éstas se transforman en frecuencias relativas, dividiendo los valores de la matriz por la frecuencia total, dando lugar a una nueva matriz que se denominará **P**.

Se obtienen las marginales de las filas *(mf)* y de las columnas *(mc)* de la matriz **P** (Tabla 5). A estos vectores se los conoce también como *Masas de fila y columna*, respectivamente.

	V_1 : *Alto*	V_1 : *Medio*	V_1 : *Bajo*	V_2 : *Alto*	V_2 : *Medio*	V_2 : *Bajo*	$\cdots$	V_p : *Alto*	V_p : *Medio*	V_p : *Bajo*
V_1 : *Alto*	$b_{1,1}$	0	0	$b_{1,4}$	$b_{1,5}$	$b_{1,6}$	$\cdots$	$b_{1,3p-2}$	$b_{1,3p-1}$	$b_{1,3p}$
V_1 : *Medio*	0	$b_{2,2}$	0	$b_{2,4}$	$b_{2,5}$	$b_{2,6}$	$\cdots$	$b_{2,3p-2}$	$b_{2,3p-1}$	$b_{2,3p}$
V_1 : *Bajo*	0	0	$b_{3,3}$	$b_{3,4}$	$b_{3,5}$	$b_{3,6}$	$\cdots$	$b_{3,3p-2}$	$b_{3,3p-1}$	$b_{3,3p}$
V_2 : *Alto*	$b_{4,1}$	$b_{4,2}$	$b_{4,3}$	$b_{4,4}$	0	0	$\cdots$	$b_{4,3p-2}$	$b_{4,3p-1}$	$b_{4,3p}$
V_2 : *Medio*	$b_{5,1}$	$b_{5,2}$	$b_{5,3}$	0	$b_{5,5}$	0	$\cdots$	$b_{5,3p-2}$	$b_{5,3p-1}$	$b_{5,3p}$
V_2 : *Bajo*	$b_{6,1}$	$b_{6,2}$	$b_{6,3}$	0	0	$b_{6,6}$	$\cdots$	$b_{6,3p-2}$	$b_{6,3p-1}$	$b_{6,3p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
V_p : *Alto*	$b_{3p-2,1}$	$b_{3p-2,2}$	$b_{3p-2,3}$	$b_{3p-2,4}$	$b_{3p-2,5}$	$b_{3p-2,6}$	$\cdots$	$b_{3p-2,3p-2}$	0	0
V_p : *Medio*	$b_{3p-1,1}$	$b_{3p-1,2}$	$b_{3p-1,3}$	$b_{3p-1,4}$	$b_{3p-1,5}$	$b_{3p-1,6}$	$\cdots$	0	$b_{3p-1,3p-1}$	0
V_p : *Bajo*	$b_{3p,1}$	$b_{3p,2}$	$b_{3p,3}$	$b_{3p,4}$	$b_{3p,5}$	$b_{3p,6}$	$\cdots$	0	0	$b_{3p,3p}$

Table 5. P: Tabla de contingencia de Burt en frecuencias relativas

V_1 : *Alto*	V_1 : *Medio*	V_1 : *Bajo*	V_2 : *Alto*	V_2 : *Medio*	V_2 : *Bajo*	$\cdots$	V_p : *Alto*	V_p : *Medio*	V_p : *Bajo*
$b_{\bullet,1}$	$b_{\bullet,2}$	$b_{\bullet,3}$	$b_{\bullet,4}$	$b_{\bullet,5}$	$b_{\bullet,6}$	$\cdots$	$b_{\bullet,3p-2}$	$b_{\bullet,3p-1}$	$b_{\bullet,3p}$

Table 6. Frecuencias marginales de las filas. (mf)

V_1 : *Alto*	V_1 : *Medio*	V_1 : *Bajo*	V_2 : *Alto*	V_2 : *Medio*	V_2 : *Bajo*	$\cdots$	V_p : *Alto*	V_p : *Medio*	V_p : *Bajo*
$b_{\bullet,1}$	$b_{\bullet,2}$	$b_{\bullet,3}$	$b_{\bullet,4}$	$b_{\bullet,5}$	$b_{\bullet,6}$	$\cdots$	$b_{\bullet,3p-2}$	$b_{\bullet,3p-1}$	$b_{\bullet,3p}$

Table 7. Frecuencias marginales de las columnas. (mc)

Se obtiene la matriz de residuos estandarizados **S**.

$$\mathbf{S} = \mathbf{D_{fila}}^{-\frac{1}{2}}(\mathbf{P} - \mathbf{mf}\ \ \mathbf{mc}')\mathbf{D_{columna}}^{-\frac{1}{2}} \tag{2}$$

donde $\mathbf{D_{fila}}$ es una matriz diagonal que contiene las masas de las filas y $\mathbf{D_{columna}}$ es una matriz diagonal que contiene las masas de las columnas.

Se aplica descomposición singular (SVD) a la matriz **S** (Ecuación 2):

$$\mathbf{S} = \mathbf{UDV}' \tag{3}$$

donde **U** y **V** son matrices ortogonales y **D** es una matriz diagonal que contiene los valores singulares.

Para encontrar las coordenadas estandarizadas se aplica lo siguiente:

$$\mathbf{X} = \mathbf{D_{fila}}^{-\frac{1}{2}}\mathbf{U} \tag{4}$$

$$\mathbf{Y} = \mathbf{D_{columna}}^{-\frac{1}{2}}\mathbf{V} \tag{5}$$

Para los fines necesarios, se utilizará las coordenadas de las columnas (Tabla 8).

2.2.2. Generalización a *k* tablas

Si se tienen *k* tablas, con la misma estructura de la tabla 3, como se visualiza en la figura 1, se aborda el enfoque del análisis factorial múltiple (MFA). Escofier and Pagès [42] indican que el MFA utiliza análisis de correspondencias múltiples cuando se trata de variables cualitativas. El

	Dim_1	Dim_2	$\cdots$	Dim_{3p}
$V_1 : Alto$	$v_1d_{1_{alto}}$	$v_1d_{1_{alto}}$	$\cdots$	$v_1d_{p_{alto}}$
$V_1 : Medio$	$v_1d_{1_{medio}}$	$v_1d_{1_{medio}}$	$\cdots$	$v_1d_{p_{medio}}$
$V_1 : Bajo$	$v_1d_{1_{bajo}}$	$v_1d_{1_{bajo}}$	$\cdots$	$v_1d_{p_{bajo}}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$V_p : Bajo$	$v_pd_{1_{bajo}}$	$v_pd_{1_{bajo}}$	$\cdots$	$v_pd_{p_{bajo}}$

Table 8. Coordenadas estandarizadas de las columnas.

procedimiento implica la realización de un MCA por cada tabla y dividirlo para su primer valor propio con la finalidad de obtener K grupos normalizados.

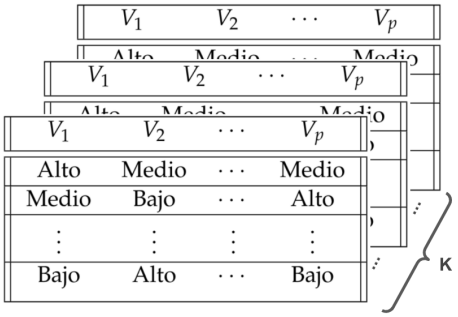

Figure 1. k tablas con el formato inicial.

La generalización a k tablas del procedimiento del MCA, se presenta en la Figura 2

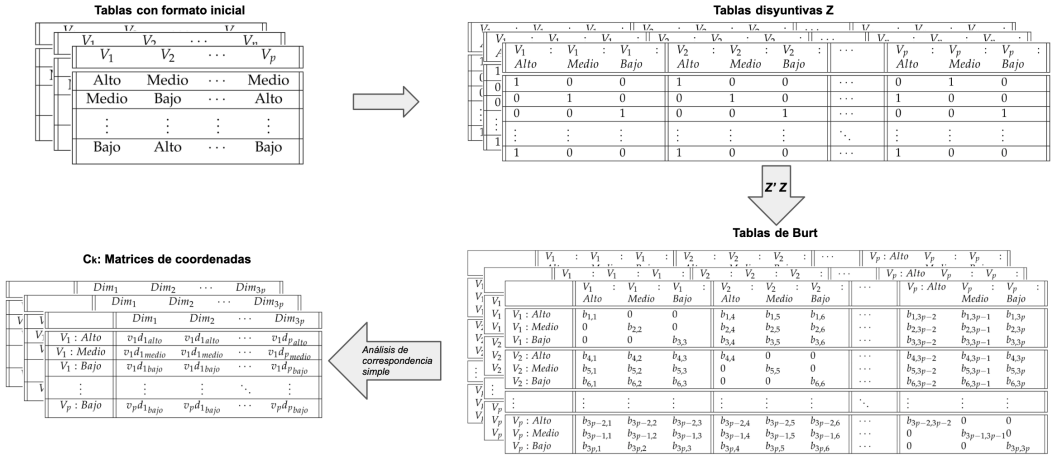

Figure 2. Procedimiento del MCA para k tablas

Se llama C a cada tabla de coordenadas. Con la finalidad de detectar la magnitud de las variables latentes, su aporte neto a las variables, se trata la matriz C con valor absoluto.
Hasta este punto se tiene un conjunto de matrices de coordenadas, cuyas filas contienen las variables observadas y las columnas, las variables latentes.

2.2.3. Normalización de tablas

Una vez que se tienen las coordenadas de las columnas, se procede a realizar la normalización, característica del procedimiento *Análisis factorial múltiple (MFA)*.

Sea λ_1^k el primer valor propio obtenido de la descomposición singular de la k-ésima tabla C. Se normaliza la tabla multiplicándola por $1/\lambda_1^k$. Con esto se obtiene la tabla $C^{'}$, que corresponde a la tabla de coordenadas normalizadas.

Individualmente, para el caso de la matriz k, se tendría la siguiente expresión.

$$\mathbf{C_k'} = \frac{1}{\lambda_k^1}\mathbf{C_k} \tag{6}$$

La expresión de la ecuación 6 aplicada a k tablas se representa en la figura 3, que muestra el esquema de preparación de las tablas, previo a la obtención de vectores de centralidad usados por el gráfico de control multivariante.

Hasta este punto se tiene un conjunto de matrices de coordenadas normalizadas, cuyas filas contienen las variables observadas y las columnas, las variables latentes.

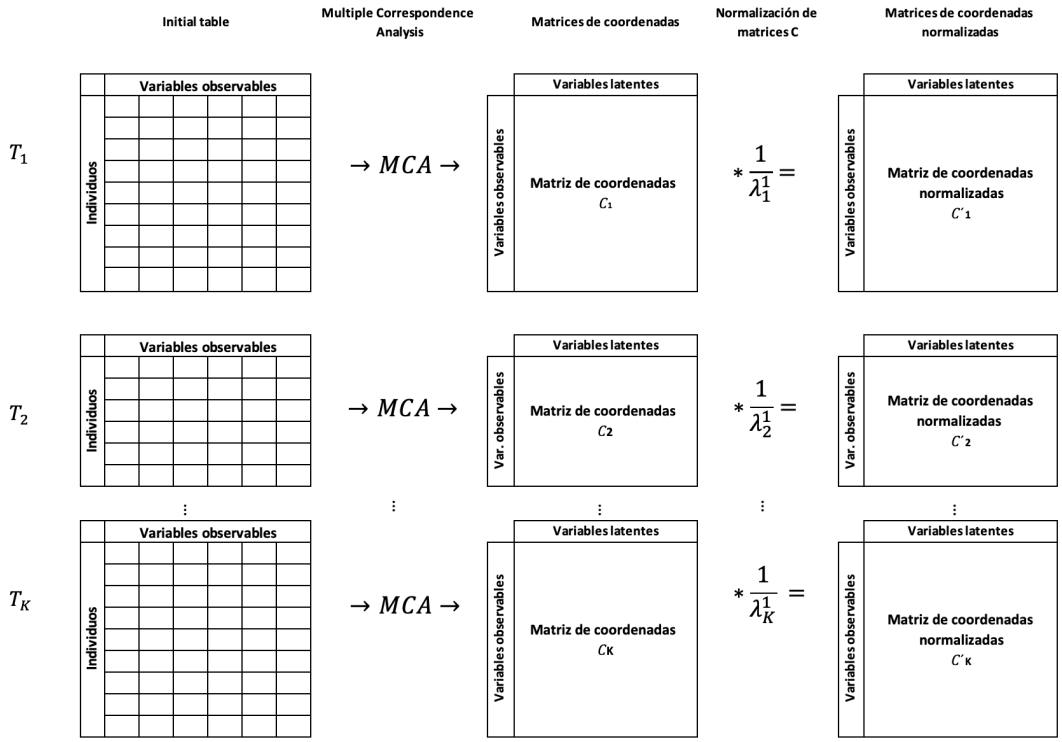

Figure 3. Esquema de preparación de las k tablas.

Aglomerando las matrices normalizadas $C^{'}$ en una sola, se tiene la matriz $\mathbb{C}^{'}$, denominada Matriz Concatenada. Esta contiene todos los elementos de las k tablas normalizadas.

$$\mathbb{C}^{'} = [\mathbf{C_1'}|\mathbf{C_2'}|,...,|\mathbf{C_K'}]^T \tag{7}$$

La normalización que realiza el MFA se encarga de ponderar las k tablas, con el objetivo de evitar alguna descompensación al momento de realizar el análisis conjunto de las tablas.

Una vez que se tiene las matrices $\mathbb{C}^{'}$ y $\mathbf{C_k'}$, se procede a obtener los vectores de mediana, tal como se muestra en la figura 4. El vector $\tilde{\mathbf{x}}_{\mathbf{C_k'}}$ explicará el comportamiento central de la tabla k y el vector $\tilde{\mathbf{x}}_{\mathbb{C}'}$ explicará el comportamiento de la matriz concatenada.

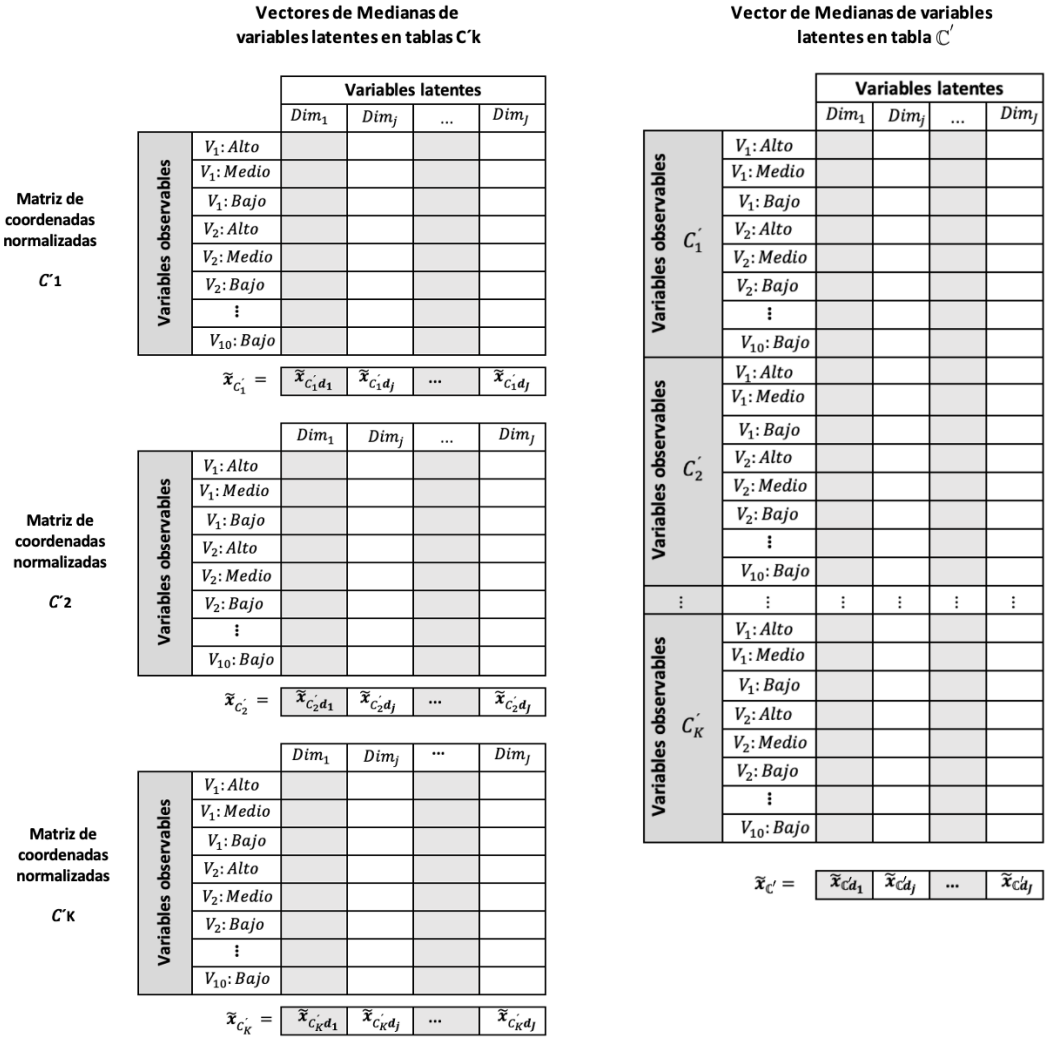

Figure 4. Esquema de obtención de vectores de medianas

2.3. *Gráfico de control T2Qv*

2.3.1. Obtención del gráfico de control

Para definir el gráfico de control T^2 Hotelling se deben tomar las siguientes consideraciones:

- La tabla $\mathbb{C}^{'}$ (Ecuación 7) se denomina Concatenada, sirve como referente para el escenario *bajo control*.
- El estadístico T^2 Hotelling normalmente se calcula con los vectores de media y considera la variabilidad de todas las tablas, incluyendo las que estén potencialmente contaminadas, es decir, las que se representarían como puntos fuera de control en el gráfico T2Qv. Esto puede generar un problema, dado que, en esta fase I se espera que la matriz concatenada capture el comportamiento "en control". Para resolver este inconveniente se suele excluir del análisis a las tablas que difieren del comportamiento normal, sin embargo, otra alternativa es adoptar conceptos de robustez, utilizando el vector de medianas en vez de el de medias, en virtud de que a las medianas no les afectan los valores atípicos.
- De la matriz concatenada $\mathbb{C}^{'}$ se obtiene $\tilde{x_0}$ (Vector de medianas de la matriz concatenada)y S_0 (Matriz de covarianzas de la matriz concatenada).
- Cada matriz $\mathbf{C_k'}$ tiene el mismo número de filas (n) y columnas (p) (individuos y variables).

- El vector de medias $\tilde{x_k}$ está atado a la tabla $\mathbf{C'_k}$, es decir, el gráfico de control estará en función de las diferencias entre las matrices $\mathbf{C'_k}$ y la matriz concatenada $\mathbb{C}^{'}$.
- Las matrices $\mathbf{C'_k}$ siguen una distribución normal multivariante con vector de centralidad $\tilde{x_k}$ y matriz de covarianzas $\mathbf{S_k}$.

El estadístico T^2 viene dado por:

$$T^2 = n(\mu_k - \mu_0)'\mathbf{\Sigma_0^{-1}}(\mu_k - \mu_0) \tag{8}$$

Tomando en cuenta las consideraciones previas, se obtiene el estadístico T^2_{med}

$$T^2_{med} = n(\tilde{x_k} - \tilde{x_0})'\mathbf{\Sigma_0^{-1}}(\tilde{x_k} - \tilde{x_0}) \tag{9}$$

Se sabe que, bajo control, el T^2 se distribuye como una Chi-cuadrado con p grados de libertad χ^2_p. En este caso se puede aplicar este principio, ya que se utiliza la matriz concatenada ($\mathbb{C}^{'}$), que representa al escenario bajo control.

Dado que este gráfico de control está basado en distancias de Mahalanobis ponderadas, sólo tiene límite de control superior. Este viene dado por la ecuación 10

$$UCL = \chi^2_{\alpha,p} \tag{10}$$

donde p es el número de dimensiones y α es la significancia predeterminada, se considera $\alpha = 0.0027$.

2.3.2. Interpretación de puntos fuera de control

El gráfico multivariante T^2 de Hotelling para variables cualitativas es capaz de señalar que el proceso salió de control, pero no permite reconocer el momento ni las causas por las que ocurrió esto. Es obvio que, más allá de reconocer el estado del proceso, interesa saber cuándo y por qué salió de control. Es importante tener en cuenta que cada punto representado en el gráfico T^2 de Hotelling representa a una tabla (muestra), constituida por un grupo de individuos (observaciones) y p variables que pueden tener muchas categorías, algunas de éstas pueden mostrar un comportamiento anómalo. Por consiguiente, es necesario analizar con detenimiento que está pasando con los datos de las tablas reportadas.

Este análisis se realiza comparando la ubicación de los puntos que representan las categorías de las variables en el MCA de la tabla concatenada y la ubicación de los puntos en los gráficos MCA de cada tabla reportada como fuera de control. Las categorías que están incidiendo en el estado fuera de control son aquellas cuya ubicación en ambas tablas comparadas muestra diferencias importantes. Para cuantificar la magnitud del comportamiento anómalo de estas categorías se calcula las distancias Chi-cuadrado entre las masas de las columnas de la tabla reportada como fuera de control y las de la tabla concatenada, tomada como referente. Mientras mayor es el valor del estadístico, mayor es su incidencia en el desplazamiento de la media del proceso que, finalmente, pueden llevarlo a un estado fuera de control.

2.4. Gráfico de diferencias Chi Cuadrado

Consiste en un gráfico interactivo de barras que representa a las distancias χ^2 entre las masas de columna de las variables de la tabla concatenada y la tabla i, que podría ser la que está fuera de control u otra que se quisiera analizar. Las barras que denotan mayor altura son las que más están incidiendo en la variación de la tendencia central del proceso y, por consiguiente, su salida de control. Para proporcionar mayor detalle, este gráfico interactivo también ofrece, mediante un gráfico circular anidado, una representación de la distribución de las categorías de la variable observada, correspondiente a la tabla i, así como un gráfico circular de la distribución de las categorías de la tabla concatenada.

De esta manera, la metodología propuesta en esta investigación permite explicar cuándo y por qué el proceso salió de control.

3. Complemento computacional

Para facilitar la difusión y aplicación del método propuesto, se ha desarrollado un paquete reproducible en R. El paquete **T2Qv** [43] realiza el análisis de control de k tablas por medio de gráficos de control multivariantes para variables cualitativas, utilizando los fundamentos del análisis de correspondencias múltiples y el análisis de factores múltiples. Los gráficos se pueden mostrar de forma plana o interactiva, de la misma manera todas las salidas se pueden mostrar en un panel interactivo de Shiny.

3.1. Disponibilidad

El paquete está disponible en el repositorio oficial de R, The Comprehensive R Archive Network (CRAN), la descarga se la puede realizar de la siguiente forma:

```
install.packages("T2Qv")
```

3.2. El paquete: T2Qv

T2Qv: Control Qualitative Variables

Covers k-table control analysis using multivariate control charts for qualitative variables using fundamentals of multiple correspondence analysis and multiple factor analysis. The graphs can be shown in a flat or interactive way, in the same way all the outputs can be shown in an interactive shiny panel.

Version:	0.1.0
Depends:	R (≥ 3.5)
Imports:	shiny, shinydashboardPlus, shinydashboard, shinycssloaders, dplyr, ca, highcharter, stringr, tables, purrr, tidyr, htmltools (≥ 0.5.1.1)
Suggests:	testthat (≥ 3.0.0)
Published:	2022-05-18
Author:	Wilson Rojas-Preciado [aut, cre], Mauricio Rojas-Campuzano [aut, ctb], Purificación Galindo-Villardón [aut, ctb], Omar Ruiz-Barzola [aut, ctb]
Maintainer:	Wilson Rojas-Preciado <wrojas at utmachala.edu.ec>
License:	MIT + file LICENSE
NeedsCompilation:	no
CRAN checks:	T2Qv results

Documentation:

Reference manual: T2Qv.pdf

Downloads:

Package source:	T2Qv_0.1.0.tar.gz
Windows binaries:	r-devel: T2Qv_0.1.0.zip, r-release: T2Qv_0.1.0.zip, r-oldrel: T2Qv_0.1.0.zip
macOS binaries:	r-release (arm64): T2Qv_0.1.0.tgz, r-oldrel (arm64): T2Qv_0.1.0.tgz, r-release (x86_64): T2Qv_0.1.0.tgz, r-oldrel (x86_64): T2Qv_0.1.0.tgz

Linking:

Please use the canonical form https://CRAN.R-project.org/package=T2Qv to link to this page.

Figure 5. Documentación del paquete T2Qv

Las funciones que contiene el paquete y su descripción se enuncian en la tabla 9.

4. Resultados

Con la intención de probar la metodología propuesta en el gráfico de control T^2 de Hotelling para variables cualitativas, se hizo un análisis con datos simulados y otro con datos reales aplicados al contexto de la educación superior. Los resultados se obtienen de la aplicación del paquete T2Qv.

Función	Descripción
T2 qualitative	Multivariate control chart T2 Hotelling applicable for qualitative variables.
MCAconcatenated	Multiple correspondence analysis applied to a concatenated table.
MCApoint	Multiple correspondence analysis applied to a specific table.
ChiSq variable	Contains Chi square distance between the column masses of the table specified in PointTable and the concatenated table. It allows to identify which mode is responsible for the anomaly in the table in which it is located.
Full Panel	A shiny panel complete with the multivariate control chart for qualitative variables, the two MCA charts and the modality distance table. Within the dashboard, arguments such as type I error and dimensionality can be modified.

Table 9. Funciones del paquete T2Qv

4.1. Resultados con datos simulados

4.1.1. Generación de datos simulados

Para este estudio se generó una base de datos simulados, a la que se denominó *Datak10Contaminated*. Consta de 10 tablas, cada una de ellas está constituida por 100 filas (observaciones) y 11 columnas, de las cuales, las 10 primeras corresponden a las variables analizadas (V1, V2, ...; V10), mismas que contienen 3 categorías (Alto, Medio y Bajo), mientras que, la columna 11, denominada *GroupLetter*, contiene el factor de clasificación de los grupos. Para su identificación, las tablas han sido denominadas con las letras del alfabeto, desde la *a* hasta la *j*. La tabla *j* tiene una distribución distinta de la que tienen las otras nueve. Las 9 primeras tablas tienen sus 10 variables con la siguiente distribución:

$$u \sim U[0,1]$$

$$t_{1,..,9} = \begin{cases} Bajo & si \quad u \leq 1/3 \\ Medio & si \quad 1/3 < u < 2/3 \\ Alto & si \quad u \geq 2/3 \end{cases}$$

La tabla 10, en todas sus 10 variables, sigue la distribución presentada a continuación:

$$u \sim U[0,1]$$

$$t_{10} = \begin{cases} Bajo & si \quad u \leq 1/5 \\ Medio & si \quad 1/5 < u < 2/6 \\ Alto & si \quad u \geq 2/6 \end{cases}$$

La base de datos se presenta en el formato establecido en la tabla 10.

Para verificar la diferencia entre las distribuciones de la tabla 10 y las demás, se calculó el promedio de las frecuencias relativas en las tres categorías, desde la tabla a hasta la i, para las 10 variables (Tabla 11), luego se calculó el promedio de las frecuencias relativas medias de las 10 variables, el resultado permite comparar la distribución de las categorías de la tabla Datak10Contaminated con la distribución teórica uniforme, como se observa en la tabla 12.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	GroupLetter
Low	Medium	Medium	High	High	High	Low	Medium	Medium	Medium	a
Low	Low	High	Low	Medium	High	High	High	Low	High	a
High	Medium	High	Low	High	Medium	Medium	High	Medium	Low	a
Medium	Medium	Low	High	Low	Medium	High	Low	Low	High	a
Low	Low	Low	High	Low	High	High	High	Medium	Medium	a
High	High	Medium	Low	High	Low	Medium	Medium	High	Low	a
High	High	Low	Low	Low	Medium	High	Medium	Medium	High	a
Medium	Medium	High	Medium	Medium	High	Medium	High	High	High	a
Low	Low	Low	Medium	High	Medium	Low	Medium	Low	Low	a
Medium	Medium	Medium	High	Low	Medium	High	Low	High	Medium	a

Table 10. Sección de la base de datos Datak10Contaminated.

Tabla	Categoría	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
a	**High**	0.29	0.25	0.36	0.38	0.38	0.35	0.36	0.29	0.33	0.37
a	**Medium**	0.36	0.49	0.34	0.34	0.31	0.41	0.38	0.28	0.38	0.31
a	**Low**	0.35	0.26	0.30	0.28	0.31	0.24	0.26	0.43	0.29	0.32
b	**High**	0.31	0.44	0.37	0.29	0.31	0.34	0.30	0.36	0.29	0.34
b	**Medium**	0.40	0.31	0.30	0.35	0.37	0.32	0.35	0.30	0.39	0.36
b	**Low**	0.29	0.25	0.33	0.36	0.32	0.34	0.35	0.34	0.32	0.30
c	**High**	0.34	0.33	0.25	0.35	0.32	0.30	0.39	0.40	0.41	0.43
c	**Medium**	0.36	0.33	0.25	0.32	0.32	0.32	0.27	0.35	0.32	0.35
c	**Low**	0.30	0.34	0.50	0.33	0.36	0.38	0.34	0.25	0.27	0.22
d	**High**	0.32	0.34	0.34	0.38	0.41	0.33	0.35	0.46	0.34	0.45
d	**Medium**	0.35	0.30	0.28	0.31	0.27	0.35	0.30	0.24	0.33	0.24
d	**Low**	0.33	0.36	0.38	0.31	0.32	0.32	0.35	0.30	0.33	0.31
e	**High**	0.32	0.32	0.36	0.26	0.36	0.31	0.29	0.28	0.32	0.41
e	**Medium**	0.34	0.40	0.34	0.40	0.38	0.37	0.27	0.37	0.32	0.23
e	**Low**	0.34	0.28	0.30	0.34	0.26	0.32	0.44	0.35	0.36	0.36
f	**High**	0.31	0.29	0.27	0.32	0.36	0.32	0.26	0.41	0.34	0.26
f	**Medium**	0.41	0.29	0.36	0.31	0.31	0.38	0.36	0.33	0.30	0.37
f	**Low**	0.28	0.42	0.37	0.37	0.33	0.30	0.38	0.26	0.36	0.37
g	**High**	0.27	0.39	0.34	0.38	0.28	0.31	0.35	0.38	0.27	0.34
g	**Medium**	0.42	0.27	0.32	0.35	0.37	0.32	0.35	0.36	0.41	0.26
g	**Low**	0.31	0.34	0.34	0.27	0.35	0.37	0.30	0.26	0.32	0.40
h	**High**	0.32	0.47	0.34	0.38	0.47	0.34	0.32	0.35	0.35	0.31
h	**Medium**	0.28	0.31	0.29	0.27	0.27	0.43	0.39	0.35	0.36	0.40
h	**Low**	0.40	0.22	0.37	0.35	0.26	0.23	0.29	0.30	0.29	0.29
i	**High**	0.32	0.42	0.29	0.30	0.26	0.28	0.38	0.38	0.36	0.36
i	**Medium**	0.35	0.34	0.29	0.33	0.47	0.38	0.25	0.29	0.33	0.31
i	**Low**	0.33	0.24	0.42	0.37	0.27	0.34	0.37	0.33	0.31	0.33
j	**High**	0.75	0.71	0.78	0.71	0.70	0.73	0.69	0.66	0.73	0.78
j	**Medium**	0.08	0.10	0.01	0.06	0.10	0.12	0.11	0.12	0.12	0.10
j	**Low**	0.17	0.19	0.21	0.23	0.20	0.15	0.20	0.22	0.15	0.12
$\bar{x}$ a: i	**High**	**0.31**	**0.37**	**0.33**	**0.33**	**0.35**	**0.32**	**0.33**	**0.37**	**0.33**	**0.36**
$\bar{x}$ a: i	**Medium**	**0.37**	**0.34**	**0.31**	**0.33**	**0.34**	**0.36**	**0.33**	**0.32**	**0.35**	**0.32**
$\bar{x}$ a: i	**Low**	**0.32**	**0.30**	**0.36**	**0.33**	**0.31**	**0.32**	**0.34**	**0.32**	**0.32**	**0.32**

Table 11. Promedio de frecuencias relativas medias en las tres categorías, desde la tabla a hasta la i. Data10Contaminated

Categorías	Teórica uniforme	Promedio Tablas a: i	Promedio Tabla j
High	0.333	0.340	0.724
Medium	0.333	0.336	0.092
Low	0.333	0.324	0.184

Table 12. Comparación de la distribución de las categorías de la tabla *Datak10Contaminated* con la distribución teórica uniforme.

Con estos datos se aplicaron pruebas Chi cuadrado de bondad de ajuste. Las hipótesis nulas consideran que la distribución de las categorías de la tabla j es igual que la distribución de cada una de las demás tablas (uniforme Low 0.333, Medium 0.333 y High 0.333). Se encontró, con un nivel de confianza del 95%, que las distribuciones de todas las variables de la tabla j mostraron diferencias estadísticamente significativas con las distribuciones de las variables de las demás tablas, con 2 grados de libertad. La tabla 13 presenta el resumen de los estadísticos de prueba respectivos. Como conclusión se ratifica que la tabla j tiene una distribución diferente de todas las demás tablas.

GroupLetter	Estadísticos	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
a	Chi-cuadrado	0.86	11.06	0.56	1.52	0.98	4.46	2.48	4.22	1.22	0.62
	p-valor	*0.651*	*0.004*	*0.756*	*0.468*	*0.613*	*0.108*	*0.289*	*0.121*	*0.543*	*0.733*
b	Chi-cuadrado	2.06	5.66	0.74	0.86	0.62	0.08	0.50	0.56	1.58	0.56
	p-valor	*0.357*	*0.059*	*0.691*	*0.651*	*0.733*	*0.961*	*0.779*	*0.756*	*0.454*	*0.756*
c	Chi-cuadrado	0.56	0.02	12.50	0.14	0.32	1.04	2.18	3.50	3.02	6.74
	p-valor	*0.756*	*0.990*	*0.002*	*0.932*	*0.852*	*0.595*	*0.336*	*0.174*	*0.221*	*0.034*
d	Chi-cuadrado	0.14	0.56	1.52	0.98	3.02	0.14	0.50	7.76	0.02	6.86
	p-valor	*0.932*	*0.756*	*0.468*	*0.613*	*0.221*	*0.932*	*0.779*	*0.021*	*0.990*	*0.032*
e	Chi-cuadrado	0.08	2.24	0.56	2.96	2.48	0.62	5.18	1.34	0.32	5.18
	p-valor	*0.961*	*0.326*	*0.756*	*0.228*	*0.289*	*0.733*	*0.075*	*0.512*	*0.852*	*0.075*
f	Chi-cuadrado	2.78	3.38	1.82	0.62	0.38	1.04	2.48	3.38	0.56	2.42
	p-valor	*0.249*	*0.185*	*0.403*	*0.733*	*0.827*	*0.595*	*0.289*	*0.185*	*0.756*	*0.298*
g	Chi-cuadrado	3.62	2.18	0.08	1.94	1.34	0.62	0.50	2.48	3.02	2.96
	p-valor	*0.164*	*0.336*	*0.961*	*0.379*	*0.512*	*0.733*	*0.779*	*0.289*	*0.221*	*0.228*
h	Chi-cuadrado	2.24	9.62	0.98	1.94	8.42	6.02	1.58	0.50	0.86	2.06
	p-valor	*0.326*	*0.008*	*0.613*	*0.379*	*0.015*	*0.049*	*0.454*	*0.779*	*0.651*	*0.357*
i	Chi-cuadrado	0.14	4.88	3.38	0.74	8.42	1.52	3.14	1.22	0.38	0.38
	p-valor	*0.932*	*0.087*	*0.185*	*0.691*	*0.015*	*0.468*	*0.208*	*0.543*	*0.827*	*0.827*
j	Chi-cuadrado	79.34	65.06	95.78	68.18	62.00	70.94	58.46	49.52	70.94	89.84
	p-valor	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*

Table 13. Estadísticos de prueba de la comparación de las distribuciones de las categorías de las 10 variables entre la tabla j y las demás, Datak10Contaminated.

Otra manera de demostrar la diferencia entre las distribuciones de las categorías en las tablas de la base de datos Datak10Contaminated es la verificación del supuesto de normalidad de los residuos. Los residuos se calculan aplicando la siguiente fórmula:

$$\hat{U} = fi_{CTx} - fi_{CTu} \tag{11}$$

, donde $\hat{U} = Residuo$, fi_{CTx} =Frecuencia relativa observada de las categorías en cada una de las tablas, fi_{CTu} = Frecuencia relativa teórica de las categorías con distribución uniforme (0.333).

Tabla	Shapiro-Wilk	gl	p-valor
a	0.965	30	0.423
b	0.967	30	0.466
c	0.960	30	0.304
d	0.939	30	0.085
e	0.987	30	0.966
f	0.955	30	0.232
g	0.954	30	0.219
h	0.970	30	0.540
i	0.976	30	0.717
j	0.757	30	0.000

Table 14. Resultados de la prueba de normalidad de residuos, Datak10Contaminated.

El análisis de normalidad de los residuos clasificados por las tablas, utilizando el estadístico de Shapiro-Wilk, determinó que en las nueve primeras tablas sí se cumplía el supuesto de normalidad, pero, en la tabla j no (p-valor = 0.000012), como se observa en la tabla 14 Queda demostrado que las categorías de la tabla j tienen una distribución diferente de las categorías de las demás tablas.

4.1.2. Aplicación del paquete T2Qv con datos simulados

El primer resultado es el gráfico del Análisis de Correspondencias Múltiples (MCA) aplicado a la tabla concatenada (Figura 6). Esta tabla ha sido tomada como referente, como escenario en control para el análisis posterior de las tablas que sean reportadas como puntos fuera de control en el gráfico T2 de Hotelling.

El MCA reporta una inercia total del 63.35%, la dimensión 1 representa al 53.64% de la información, mientras que la dimensión 2, al 9.71%. Los puntos del gráfico representan a las observaciones de cada una de las 10 variables en sus tres niveles: alto, medio y bajo. En esta figura, todas las observaciones que corresponden al nivel alto se ubican a la izquierda en el eje de las X; de las 10 observaciones correspondientes al nivel medio, 8 se situaron en el cuarto cuadrante y las dos restantes en el cuadrante 1, es decir, todas las observaciones de este nivel estuvieron a la derecha en el eje de las X. Finalmente, de los 10 puntos que representan al nivel bajo, 8 están ubicados en el cuadrante 1.

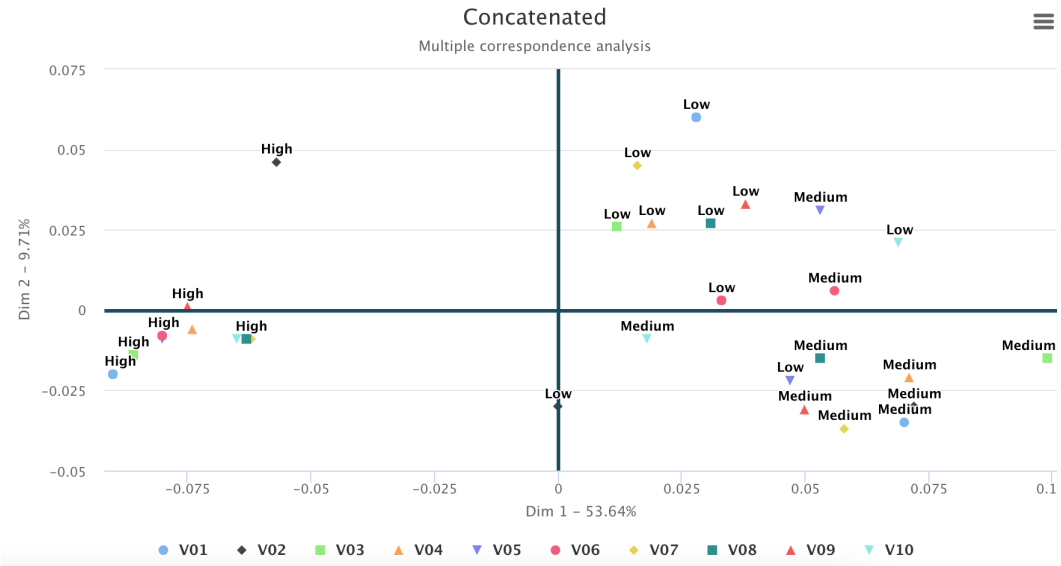

Figure 6. Análisis de correspondencias múltiples aplicado a la tabla consenso.

Otro resultado es el Análisis de Correspondencias Múltiples aplicado a una tabla específica. En este punto, uno de los argumentos que se debe tener en cuenta es la selección de la tabla con la que se realizará el análisis.

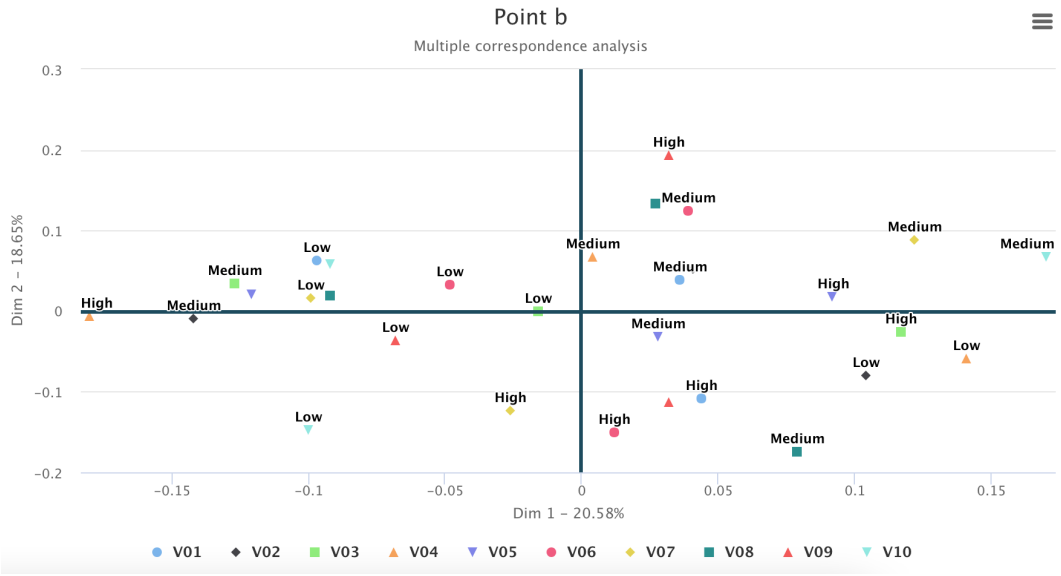

Figure 7. Análisis de correspondencias múltiples aplicado a la tabla b.

La figura 7 representa el gráfico del MCA de la tabla b. Este gráfico, en sus dos dimensiones, representa al 39.23% de la información. Es notorio que las observaciones en sus niveles alto, medio y bajo están distribuidas de forma aleatoria en todos los cuadrantes del gráfico, no se puede precisar un patrón específico de agrupación. Esto mismo se puede decir de los puntos representados en cualquiera de las otras tablas porque comparten la misma distribución, exceptuando la tabla j, que fue diseñada con una distribución diferente. No obstante, el uso del MCA de las figuras 7 y 6 todavía no permite detectar si el proceso está o no en control. La identificación de puntos fuera de control se puede realizar mediante la representación gráfica del estadístico T2 de Hotelling, como se observa en la figura 8.

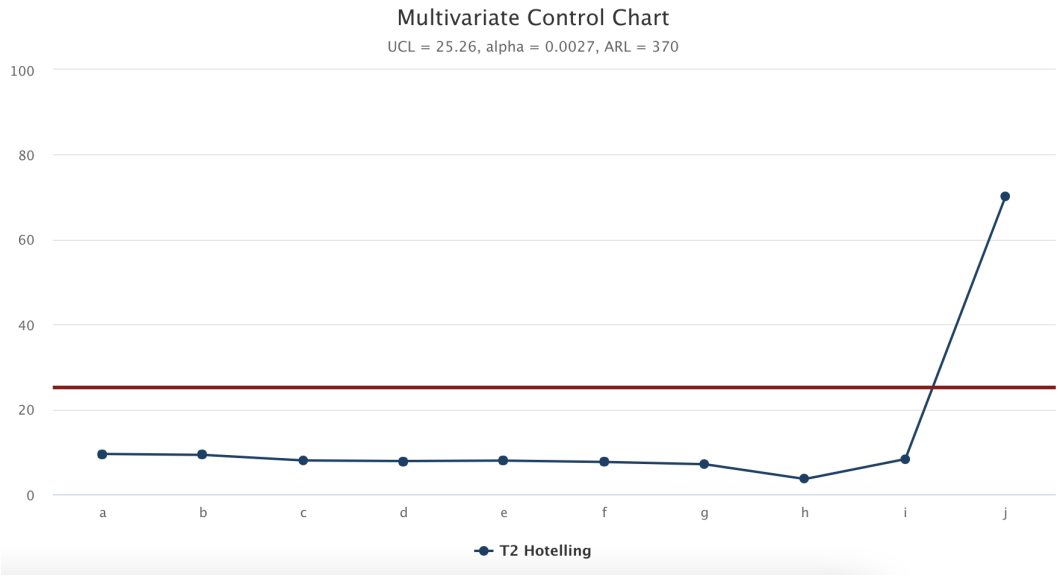

Figure 8. Gráfico de control multivariante T2 Hotelling aplicable a variables cualitativas, Datak10Contaminated.

La figura 8 presenta un gráfico de control elaborado con el estadístico T2 de Hotelling, aplicado a la detección de anomalías en cualquiera de las k tablas analizadas. Cada una de las tablas está representada por los puntos en el gráfico. Se observa una línea horizontal que representa al límite

de control superior (UCL). El límite de control inferior (LCL) es igual a cero. Dado que el análisis de sensibilidad determinó que este gráfico de control tiene un mejor rendimiento cuando trabaja con un número alto de dimensiones, se ha recomendado que este sea p-1, donde p es el número de dimensiones inicial, que es equivalente a la cantidad de variables de la base de datos, sin contar a la variable GroupLetter que sólo sirve como factor de clasificación de las tablas. Se observa que el punto que representa a la tabla j se ubica por encima del límite de control superior, lo que quiere decir que se lo ha identificado como un valor fuera de control. Por consiguiente, es necesario analizar con detenimiento qué está pasando con los datos de la tabla reportada, comparándolos con los de la tabla consenso, a fin de identificar las causas de la variación y tomar las acciones pertinentes. Para hacer un análisis del punto fuera de control se realiza un gráfico del MCA de la tabla j y se lo compara con el gráfico similar de la tabla consenso, como se presenta en la figura 9.

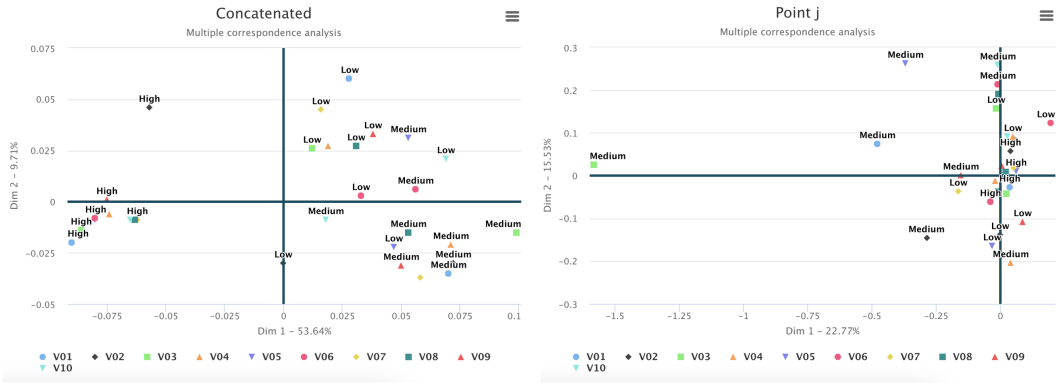

Figure 9. Gráfico de control multivariante T2 Hotelling aplicable a variables cualitativas, Datak10Contaminated.

La figura 9 presenta la distribución de las observaciones de las tablas consenso y j mediante gráficos del MCA. El gráfico de la tabla consenso, que sirve de referente en control, ya se analizó en la figura 4; el de la tabla j muestra una tendencia de los puntos que con valores medios a ubicarse al lado izquierdo, bastante alejados de los demás que confluyen hacia el centro del eje de las X. Especial atención merece la variable 3, que registra una observación para el nivel medio con el valor más alejado del grupo. Al comparar los gráficos es evidente que la distribución de los datos en el gráfico de la tabla j es diferente de las distribuciones de las demás tablas, y en especial, es diferente de la distribución de los datos en el gráfico de la tabla consenso, lo que explica por qué el punto j ha sido identificado como fuera de control en el gráfico T2 de Hotelling. Esta diferencia se explica en la tabla 15, que muestra la distancia Chi cuadrado entre las observaciones de la tabla consenso y la tabla j.

Variables	ChiSq
V1	0.06968
V2	0.05010
V3	0.07601
V4	0.04982
V5	0.05205
V6	0.05603
V7	0.03713
V8	0.03702
V9	0.04395
V10	0.06179

Table 15. Distancia Chi cuadrado entre las masas de columna de la tabla k y la consenso, Datak10Contaminated.

El comportamiento de estas variables en la tabla j provoca el desplazamiento de la tendencia central del proceso que, al final, lo lleva a un estado fuera de control. Otra manera de visualizar esta información es a través de un gráfico de barras (figura 10).

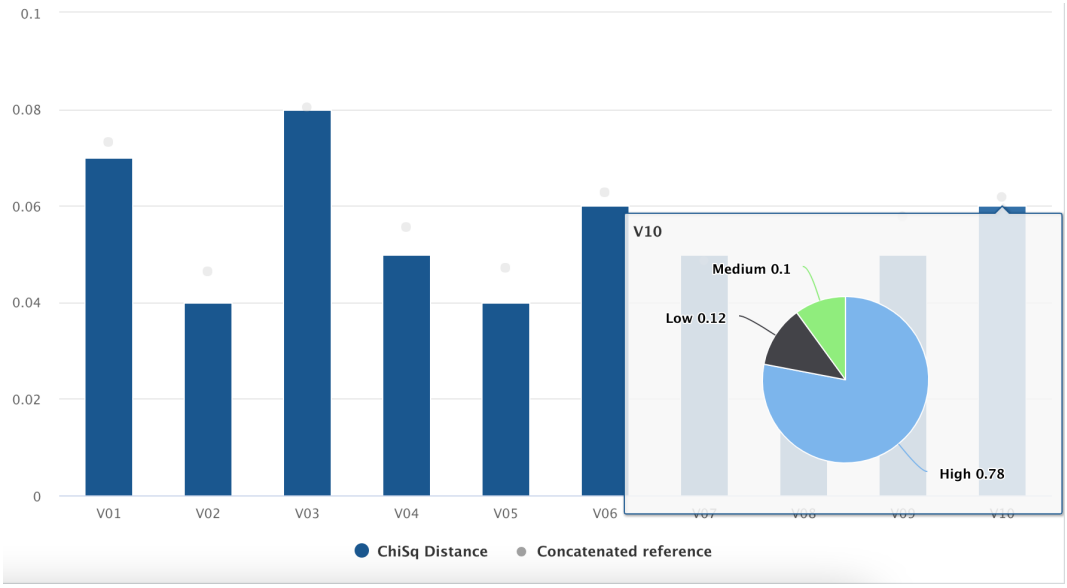

Figure 10. Distancia Chi cuadrado entre las masas de la tabla consenso y las k tablas, Datak10Contaminated

El gráfico de barras de la figura 10, expresa también la distancia χ^2 entre las masas de la tabla consenso y las de las k tablas de la base de datos Datak10Contaminated. Además, la interactividad de este gráfico facilita la observación de la distribución de las categorías de las variables de la tabla analizada, en este caso la j, y su comparación con la distribución de las categorías de las variables en la tabla consenso.

5. Discusión

En el SPC para variables cualitativas todavía no son muchas las propuestas publicadas. Las diferencias entre procedimientos para la determinación de los estadísticos y los gráficos de control en este campo hacen difícil su comparación.

El gráfico de control T2Qv, que se presenta en este artículo, aplica un MCA, técnica de análisis multivariante que identifica estructuras latentes que subyacen en el conjunto de datos cualitativos y que involucra una reducción de dimensiones, en consecuencia, desde el comienzo se requiere una tabla de datos con p variables ($p > 3$) dicotómicas o politómicas. Se debe recordar que el análisis de sensibilidad determinó que esta propuesta tiene un buen rendimiento cuando trabaja con altas dimensiones y que a bajas dimensiones pierde estabilidad. En varios estudios revisados, los casos de aplicación analizan sólo dos o tres variables, lo que conduciría a la aplicación de un análisis de correspondencias simple, no múltiple. En consecuencia, estos casos no podrían ser tratados con el T2Qv.

Como ejemplos se señala la Combinación lineal óptima de variables Poisson para el SPC multivariados, de Epprecht *et al.* [26] cuyo caso de aplicación registrado en su publicación analiza dos variables relacionadas con el conteo de defectos en la producción de jarrones de cerámica. El gráfico GMDS de Ali and Aslam [32] fue ejemplificado con un conjunto de datos de telecomunicaciones, tomado de Jiang *et al.* [44], que consta de sólo dos variables. El gráfico de control multivariante, desarrollado por Fernández *et al.* [31]), para p características de calidad de atributos correlacionadas, que aplica teoría difusa, hace un análisis de dos tablas tomadas de publicaciones de Taleb [30] y Taleb

et al. [29]), la primera con tres variables relacionadas con la comida congelada, y la segunda, con tres variables sobre la producción de porcelana.

Otra de las características del gráfico T2Qv es que cada muestra es un grupo constituido por un conjunto de individuos. El ejemplo de datos simulados *Datak10Contaminated* incluye un conjunto de 10 tablas y 11 variables, cada tabla es una muestra, está formada por 100 observaciones y aparece representada como un punto en el gráfico T^2 de Hotelling; el ejemplo aplicado al contexto educativo hace referencia a la base de datos *Estudiantes 2019_2020*, conformada por 43191 observaciones y 17 variables cualitativas agrupadas en 4 periodos académicos, estos periodos constituyen las tablas (muestras) que se representan como puntos en el gráfico. En publicaciones de varios autores se puede constatar que en sus ejemplos de aplicación se analiza una sola tabla, de dimensiones n (filas) x p (variables), donde cada n_i fila es una muestra.

Por ejemplo, el gráfico de control MNP, de Lu [27] contiene en su artículo una tabla de datos simulados de 30 muestras, donde cada una de ellas es un único individuo (objeto) que registra el conteo de defectos para tres características de la calidad. Asimismo, la ejemplificación que Chiu and Kuo [23] presentaron de su gráfico de control *MP* se hizo con una tabla de datos simulados de 26 muestras, donde cada muestra representa a un individuo al que se le registra el D número de defectos o no conformidades asociadas a tres características de calidad.

En el gráfico de control T2Qv que se presenta en este artículo, cada uno de los individuos (filas) que conforman las diferentes nuestras pueden tener distintas configuraciones en función de las categorías de las variables. En base de datos *Estudiantes_2019_2020*, por ejemplo, el primer individuo de la lista es una mujer que estudia la carrera de Acuicultura en la Facultad de Ciencias Agropecuarias, su edad está entre 18 y 30 años, no presenta discapacidad, se autodeclaró mestiza; vive en una zona urbana, su padre tiene un nivel de formación de Educación Básica, su madre también; en su hogar viven 5 o más personas, sus estudios secundarios los realizó en un colegio particular, el ingreso total de su hogar se clasifica como de Rango 2, no registró el origen de los recursos económicos para sus estudios, no tuvo necesidad de acudir a segunda ni tercera matrícula y sí terminó su periodo académico. Otros estudiantes de esta misma tabla, o de las otras tres, tendrán diferentes características, hay que considerar que en total son 43191 individuos.

Por el contrario, otros autores que han investigado sobre gráficos de control multivariante para datos de atributos, aunque en su análisis consideran varias características de calidad, al final clasifican a cada individuo por una sola de las variables analizadas. Es el caso de Mukhopadhyay [28], cuya propuesta se demuestra con un caso de aplicación que controla 7 características de calidad en 24 muestras cuyo tamaño varía entre 20 y 404 individuos. Las variables responden a 6 tipos de defectos de la pintura en la cubierta de ventiladores de techo: cobertura deficiente, desbordamiento, defecto de empanada, burbujas, defectos de pintura, defectos de pulido. La séptima característica es la ausencia de defectos. A cada individuo se lo clasifica por su defecto más predominante, por consiguiente, en su registro sólo aparece un tipo de defecto o ausencia de defectos, lo que resulta en una pérdida de información sobre el efecto combinado de las variables sobre el proceso.

6. Conclusiones

En este artículo se ha presentado el gráfico de control T2Qv, un técnica de control estadístico de procesos multivariantes que realiza un análisis de los datos cualitativos a través del Análisis de correspondencias múltiple, cuyas coordenadas se someten a un proceso de normalización propio del Análisis Factorial Múltiple, para luego representarlos mediante el gráfico T^2 de Hotelling.

Esta propuesta genera un gráfico del MCA de la tabla concatenada, que sirve de referente para comparar otros gráficos del MCA de las tablas que hayan sido identificadas como puntos fuera de control en el gráfico de Hotelling. Allí se puede verificar qué categorías de las variables han tenido variaciones en su ubicación en ambos gráficos, que pueden estar provocando cambios en la media del proceso y ocasionando el estado de fuera de control.

Para facilitar la interpretación del comportamiento de las variables se realiza un análisis de la distancia

Chi cuadrado entre las categorías de la tabla concatenada y de las tablas reportadas como fuera de control. Para este análisis se puede utilizar una tabla que reporta los valores del estadístico Chi cuadrado y los p-valores que determinan significancia estadística en tres niveles: 0.05 $(*)$, 0.01$(**)$ y 0.001$(***)$. También se puede representar este análisis mediante un gráfico de barras que incluye límites asociados a los niveles de significancia estadística establecidos.
El análisis de sensibilidad determinó que el gráfico de control T2Qv tiene un buen rendimiento cuando trabaja con altas dimensiones, pero, que pierde estabilidad a bajas dimensiones. Para facilitar la difusión y aplicación del método propuesto, se ha desarrollado un paquete estadístico computacional reproducible en R, denominado T2Qv y disponible en GitHub, que permite visualizar los resultados de forma plana o interactiva, además, presenta un panel Shiny que contiene todas las funciones integradas en un mismo espacio.

En el SPC para variables cualitativas todavía no son muchas las propuestas publicadas. Las diferencias entre procedimientos para la determinación de los estadísticos y los gráficos de control en este campo hacen difícil su comparación.

Appendix A

Appendix A.1

Appendix B

References

1. Gutiérrez, H.; de la Vara Salazar, R. *Control estadístico de la calidad y seis sigma*; Vol. 3, McGraw Hill Education, 2013; p. 152 – 253.
2. Ramos, M. Una alternativa a los métodos clásicos de control de procesos basada en coordenadas paralelas, métodos Biplot y Statis. PhD thesis, 2017.
3. Li, J.; Tsung, F.; Zou, C. Directional control schemes for multivariate categorical processes. *Journal of Quality Technology* **2012**, *44*, 136–154.
4. Hotelling, H. Multivariate quality control. Techniques of statistical analysis. *McGraw-Hill, New York* **1947**.
5. Lowry, C.A.; Woodall, W.H.; Champ, C.W.; Rigdon, S.E. A multivariate exponentially weighted moving average control chart. *Technometrics* **1992**, *34*, 46–53.
6. Crosier, R.B. Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. *Technometrics* **1988**, *30*, 291–303.
7. APARISI, F. Hotelling's T2 control chart with adaptive sample sizes. *International Journal of Production Research* **1996**, *34*, 2853–2862, [https://doi.org/10.1080/00207549608905062]. doi:10.1080/00207549608905062.
8. Aparisi, F.; Haro, C.L. Hotelling's T2 control chart with variable sampling intervals. *International Journal of Production Research* **2001**, *39*, 3127–3140, [https://doi.org/10.1080/00207540110054597]. doi:10.1080/00207540110054597.
9. Faraz, A.; Parsian, A. Hotelling's T2 control chart with double warning lines. *Statistical Papers* **2006**, *47*, 569–593. doi:10.1007/s00362-006-0307-x.
10. Shabbak, A.; Midi, H. An improvement of the hotelling statistic in monitoring multivariate quality characteristics. *Mathematical Problems in Engineering* **2012**, *2012*.
11. Kim, S.B.; Jitpitaklert, W.; Park, S.K.; Hwang, S.J. Data mining model-based control charts for multivariate and autocorrelated processes. *Expert Systems with Applications* **2012**, *39*, 2073–2081.
12. Ruiz-Barzola, O. Gráficos de Control de Calidad Multivariantes con Dimension Variable. PhD thesis, Universitat Politécnica de Valéncia, 2013.
13. Yeong, W.C.; Khoo, M.B.C.; Teoh, W.L.; Castagliola, P. A control chart for the multivariate coefficient of variation. *Quality and Reliability Engineering International* **2016**, *32*, 1213–1225.

14. Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert Systems with Applications* **2014**, *41*, 1701–1707. doi:10.1016/j.eswa.2013.08.068.
15. Ahsan, M.; Mashuri, M.; Kuswanto, H.; Prastyo, D.D.; Khusna, H. Multivariate control chart based on PCA mix for variable and attribute quality characteristics. *Production & Manufacturing Research* **2018**, *6*, 364–384, [https://doi.org/10.1080/21693277.2018.1517055]. doi:10.1080/21693277.2018.1517055.
16. Liu, Y.; Liu, Y.; Jung, U. Nonparametric multivariate control chart based on density-sensitive novelty weight for non-normal processes. *Quality Technology & Quantitative Management* **2020**, *17*, 203–215.
17. YILMAZ, H.; Yanik, S. Design of Demerit Control Charts with Fuzzy c-Means Clustering and an Application in Textile Sector. *Textile and Apparel* **2020**, *30*, 117–125.
18. Farokhnia, M.; Niaki, S.T.A. Principal component analysis-based control charts using support vector machines for multivariate non-normal distributions. *Communications in Statistics - Simulation and Computation* **2020**, *49*, 1815–1838, [https://doi.org/10.1080/03610918.2018.1506032]. doi:10.1080/03610918.2018.1506032.
19. Xue, L.; Qiu, P. A nonparametric CUSUM chart for monitoring multivariate serially correlated processes. *Journal of Quality Technology* **2020**, pp. 1–14.
20. Ahsan, M.; Mashuri, M.; Wibawati.; Khusna, H.; Lee, M.H. Multivariate Control Chart Based on Kernel PCA for Monitoring Mixed Variable and Attribute Quality Characteristics. *Symmetry* **2020**, *12*. doi:10.3390/sym12111838.
21. Ahsan, M.; Mashuri, M.; Kuswanto, H.; Prastyo, D.D.; Khusna, H. Outlier detection using PCA mix based T2 control chart for continuous and categorical data. *Communications in Statistics - Simulation and Computation* **2021**, *50*, 1496–1523, [https://doi.org/10.1080/03610918.2019.1586921]. doi:10.1080/03610918.2019.1586921.
22. Holgate, P. Estimation for the bivariate Poisson distribution. *Biometrika* **1964**, *51*, 241–287.
23. Chiu, J.E.; Kuo, T.I. Attribute control chart for multivariate Poisson distribution. *Communications in Statistics-Theory and Methods* **2007**, *37*, 146–158.
24. Lee, L.H.; Costa, A.F.B. Control charts for individual observations of a bivariate Poisson process. *The International Journal of Advanced Manufacturing Technology* **2009**, *43*, 744–755.
25. Laungrungrong, B.; M, C.B.; Montgomery, D.C. EWMA control charts for multivariate Poisson-distributed data. *International Journal of Quality Engineering and Technology* **2011**, *2*, 185–211.
26. Epprecht, E.K.; Aparisi, F.; García-Bustos, S. Optimal linear combination of Poisson variables for multivariate statistical process control. *Computers & operations research* **2013**, *40*, 3021–3032.
27. Lu, X. Control chart for multivariate attribute processes. *International Journal of Production Research* **1998**, *36*, 3477–3489.
28. Mukhopadhyay, A.R. Multivariate attribute control chart using Mahalanobis D 2 statistic. *Journal of Applied Statistics* **2008**, *35*, 421–429.
29. Taleb, H.; Limam, M.; Hirota, K. Multivariate fuzzy multinomial control charts. *Quality Technology & Quantitative Management* **2006**, *3*, 437–453.
30. Taleb, H. Control charts applications for multivariate attribute processes. *Computers & Industrial Engineering* **2009**, *56*, 399–410.
31. Fernández, M.N.P.; García, A.C.; Barzola, O.R. Multivariate multinomial T 2 control chart using fuzzy approach. *International Journal of Production Research* **2015**, *53*, 2225–2238.
32. Ali, M.R.; Aslam, M. Design of control charts for multivariate Poisson distribution using generalized multiple dependent state sampling. *Quality Technology & Quantitative Management* **2019**, *16*, 629–650.
33. Saltos Segura, G.; Flores Sánchez, M.; Horna Huaraca, L.; Morales Quinga, K. NEW METHODOLOGIES APPLIED TO MULTIVARIATE MONITORING OF STUDENT PERFORMANCE USING CONTROL CHARTS AND THRESHOLD SYSTEMS. *Perfiles* **2020**, *1*, 68–74.
34. López, C.P. *Técnicas de análisis multivariante de datos*; Pearson Educación, 2004.
35. Pearson, K. LIII.On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6* **1901**, *2*, 417 – 441. doi:10.1080/14786440109462720.
36. Hotelling, H. Analysis of a complex of statistical variables into principal components. **1933**. *24*, 417 – 441. doi:10.1037/h0071325.
37. Ch, S.; others. General intelligence objectively determined and measured. *American Journal of Psychology* **1904**, *15*, 201–293.

38. Thurstone, L.L. Multiple-factor analysis; a development and expansion of The Vectors of Mind. **1947**.
39. Kaiser, H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200.
40. Benzecri., J. *OL'analyse des correspondances. En L'Analyse des Données: Leçons sur L'analyse Factorielle et la Reconnaissance des Formes et Travaux*; Paris - 1973, 1973.
41. Michailidis, G.; Leeuw, J.D. The Gifi system of descriptive multivariate analysis. *Statistical Science* **1998**, pp. 307–336.
42. Escofier, B.; Pagès, J. Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis* **1994**, *18*, 121–140. doi:https://doi.org/10.1016/0167-9473(94)90135-X.
43. Rojas-Preciado, W.; Rojas-Campuzano, M.; Galindo-Villardón, P.; Ruiz-Barzola, O. *T2Qv: Control Qualitative Variables*. R package version 0.1.0.
44. Jiang, W.; Au, S.; Tsui, K.L.; Xie, M. Process monitoring with univariate and multivariate c-charts. *Technical Report, the Logistics Institute, Georgia Tech, and the Logistics Institute-Asia Pacific* **2002**.